From Efficiency to Equity: Measuring Fairness in Preference Learning

Shreeyash Gowaikar[1], **Hugo Berard**[2], **Rashid Mushkani**[2], **Shin Koseki**[2]

[1]Microsoft Research India
[2]University of Montréal
shreeyashgo@gmail.com, hugo.berard@umontreal.ca, rashid.ahmad.mushkani@umontreal.ca, shin.koseki@umontreal.ca

Abstract

As AI systems, particularly generative models, increasingly influence decision-making, ensuring that they are able to fairly represent diverse human preferences becomes crucial. This paper introduces a novel framework for evaluating epistemic fairness in preference learning models inspired by economic theories of inequality and Rawlsian justice. We propose metrics adapted from the Gini coefficient, the Atkinson index, and the Kuznets ratio to quantify fairness in these models. We validate our approach on three different tasks where we have access to individual annotator preferences and compare how well each user is represented by a preference learning model. Our analysis reveals variations in model performance between users, highlighting potential epistemic injustices. We explore pre-processing and in-processing techniques to mitigate these inequalities, demonstrating a complex relationship between model efficiency and fairness. This work contributes to AI ethics by providing a framework for evaluating and improving epistemic fairness in preference learning models, offering insights for developing more inclusive AI systems in contexts where diverse human preferences are crucial.

Introduction

The rapid advancement of generative artificial intelligence (AI) has brought unprecedented capabilities in natural language processing and content generation. However, these developments have also raised significant concerns about the potential for these systems to perpetuate or amplify epistemic injustice (Kay, Kasirzadeh, and Mohamed 2024). Epistemic injustice, a concept introduced by Miranda Fricker (Fricker 2007), refers to the wrongs done to individuals in their capacity as knowers. In the context of generative AI, this manifests itself as misrepresentation or misunderstanding of the views of minorities and marginalized groups, which can subject them to epistemic violence by denying their own subjective experience.

Generative AI systems, such as large language models, risk producing epistemic injustices by embedding biases through their training data and amplifying voices from majority groups while silencing voices from minorities. This imposition of a singular framework on diverse perspectives can fail to recognize the multitude of views and experiences that exist throughout the world. The challenge, therefore, is to develop generative AI systems that can fairly represent all existing views and perspectives, acknowledging and respecting the diversity of human experiences. To address these alignment challenges, researchers have turned to techniques such as Reinforcement Learning with Human Feedback (RLHF) (Ziegler et al. 2019). RLHF typically involves a multistep process: first, gathering a dataset where human annotators indicate their preferences among a set of AI-generated options; second, training a reward model on this dataset to predict which options were preferred; and finally, using this reward model to fine-tune the generative model to align it with human preference. Although this approach can improve alignment between AI outputs and human values, it raises new questions about epistemic justice.

The concept of the "tyranny of the majority," long recognized in political theory (Mill 1859), becomes relevant in this context. If the reward model is biased towards certain groups, it may capture the preferences of dominant groups at the expense of marginalized voices. This scenario could lead to a digital manifestation of majority rule, where the opinions and values of the numerical majority consistently overshadow those of minority groups in AI-generated content. To mitigate this risk, we argue that reward models should be trained on a diverse set of preferences from annotators who are as representative of the global population as possible. However, even with a diverse dataset, the underrepresentation of minorities may still lead to their perspectives being overlooked in the final model. Therefore, it is crucial to develop methods to measure and address this form of epistemic injustice within reward models.

In this paper, we propose novel metrics inspired by economic literature on inequality and fair allocation to quantify to which extent a reward model equally captures the preferences of every users. We demonstrate the application of these metrics on two preference learning tasks, revealing that epistemic injustice can persist even in models with high overall accuracy. Our findings underscore the importance of looking beyond aggregate performance metrics to ensure equitable representation of diverse perspectives.

Furthermore, our research highlights a critical gap in the field: the scarcity of datasets containing individual annotations that would allow for a more nuanced analysis of pref-

erence distribution across different groups. We advocate for the creation and public release of such datasets, which are essential to advance research on epistemic justice in AI and to develop more inclusive generative models. By addressing these challenges, we aim to contribute to the development of generative AI systems that not only perform well on standard metrics but also uphold principles of epistemic justice, ensuring that the diversity of human knowledge and experience is respected and accurately represented in AI-generated content.

Related Work

The intersection of fairness, epistemic justice, and AI has gathered significant attention in recent years, particularly in the context of classification and regression tasks. However, less attention has been paid to fairness and epistemic considerations in more complex AI systems such as generative models and preference learning algorithms.

Fairness in Classification and Regression In traditional machine learning tasks, fairness metrics often focus on equalizing outcomes across different groups. (Hardt et al. 2016) introduced the concept of Equal Opportunity, which aims to equalize true positive rates between protected and unprotected groups. Other measures include equalized odds, ensuring equal probability of positive outcomes across classes, and equal accuracy, which balances performance across groups (Cotter et al. 2019). These metrics, while valuable, primarily address fairness in standard classification tasks, where the goal is to ensure that the algorithm's output does not depend on sensitive attributes. However, in the context of Reinforcement Learning from Human Feedback (RLHF) and preference learning, the concept of fairness requires a different approach. In these scenarios, we acknowledge that different groups may have varying preferences, and thus, the output of the reward model should rightfully depend on the user. Our notion of fairness in this context focuses on ensuring that the model's accuracy does not vary significantly across different groups and individual users, and thus is able to capture those diverse preferences.

Fairness in Preference Learning and Ranking While works like (Narasimhan et al. 2020) and (Beutel et al. 2019) have proposed fairness metrics for ranking tasks, their focus primarily remains on ensuring fair treatment of the items being ranked. (Sanyal, Hu, and Yang 2022) take a step further by discussing the equity of subgroup populations and exploring the trade-off between this equity and overall accuracy. However, there remains a crucial distinction between these approaches and ours. Standard fairness approaches in ranking typically aim to ensure that items from different groups (e.g., protected vs. unprotected) have equal opportunities to be ranked highly. This focus on the fairness of outcomes for ranked items is important but does not address the full spectrum of fairness concerns in preference learning scenarios. Our approach, in contrast, shifts the focus to the equity of the participants providing the rankings or expressing preferences. We argue that in preference learning and RLHF contexts, it's crucial to ensure that the model's ability to capture and represent preferences is equitable across all participants, regardless of their group membership. This means that while the content of preferences may vary across groups (which is expected and acceptable), the accuracy with which these preferences are captured and represented by the model should be consistent across all participants. This distinction is particularly important in scenarios where diverse viewpoints and experiences are critical, such as in collaborative decision-making or in developing AI systems that need to be responsive to a wide range of user preferences. By focusing on the equity of participants rather than just the ranked items, we aim to develop models that are truly representative of diverse human preferences and experiences. This perspective reveals a critical gap in the literature: most current approaches treat participants as interchangeable, assuming a universal preference or aggregating scores without considering individual differences (Sawyer, Cole, and Cole 1976; Bakker et al. 2022). This "One-Truth" fallacy (Aroyo and Welty 2015) fails to account for the diversity of human preferences, which can stem from personal, environmental, and socio-demographic factors (Sandri et al. 2023).

Towards Diverse Preference Representation Recent work has begun to acknowledge the importance of diverse human preferences in AI alignment and preference learning. Approaches include developing multiple reward models (Chakraborty et al. 2024; Bakker et al. 2022), multi-policy strategies (Ramé et al. 2023), and consensus-based ranking (Kovač et al. 2023). However, there remains a lack of consistency in how diversity and model performance are measured and evaluated. Most studies rely on basic classification metrics like F1 score and accuracy across users (Sandri et al. 2023; Yu et al. 2023; Zeng et al. 2024) or average reward values (Ramé et al. 2023; Bakker et al. 2022). While these metrics provide some insight, they fail to capture the nuanced ways in which AI systems might perpetuate epistemic injustice by misrepresenting or undermining the subjective experiences of marginalized groups.

Background

This section introduces a mathematical framework for preference learning that allows us to quantify both the overall performance of a model and its fairness across diverse users.

Problem Definition Consider a set of k users $\mathcal{U} = \{1, ..., k\}$ and a dataset $\mathcal{D} = \{(x_i, x_i', s_i, u_i)\}_{i=1}^{n}$ composed of n pairwise comparisons. Each entry in the dataset represents a comparison where user u_i provided a score $s_i \in \mathcal{S}$. The score can be either a categorical variable (i.e., $\mathcal{S} = \{0, 1\}$ indicating which option was preferred or a real value (i.e., $\mathcal{S} = \mathbb{R}$), where negative scores indicate preference for x_i, and positive scores preference for x_i', and the magnitude of s_i reflects the strength of the preference. Our goal is to learn a model $f : \mathcal{X} \times \mathcal{X} \to \mathcal{Y}$ that can score any pair $(x, x') \in \mathcal{X} \times \mathcal{X}$. We define the error of the model as:

$$\mathcal{E}(f) = \mathbb{E}\left[\ell\left(f(x_i, x_i'), s_i\right)\right]$$

where ℓ is a loss function that computes the discrepancy between the model outputs and the ground truth score. The choice of loss function depends on the nature of the scores:

1. For real-valued scores, we can use the squared error:

$$\ell\left(f(x_i, x_i'), s_i\right) = \left(f(x_i, x_i') - s_i\right)^2$$

2. For binary scores, where f predicts the probability that x_i is preferred over x_i', we can use the binary cross-entropy (BCE) loss:

$$\begin{aligned}\ell\left(f(x_i, x_i'), s_i\right) = s_i \log\left(f(x_i, x_i')\right) \\ + (1 - s_i) \log\left(1 - f(x_i, x_i')\right)\end{aligned}$$

3. Alternatively, we can use the 0-1 loss:

$$\ell\left(f(x_i, x_i'), s_i\right) = \begin{cases} 0 & \text{if } s_i = f(x_i, x_i') \\ 1 & \text{else} \end{cases}$$

To evaluate the model's performance for individual users, we define the user-specific error for user u:

$$\mathcal{E}_u(f) = \mathbb{E}\left[\ell\left(f(x_i, x_i'), s_i\right) | u\right]$$

Drawing inspiration from game theory literature on fair allocation, we introduce two key concepts that will help us evaluate both the overall performance and the fairness of our preference learning models.

Definition 1 (Efficiency) *This efficiency measures the overall performance of the model across all users. For a model f, it is the mean of the errors across all users:*

$$\bar{\mathcal{E}}(f) = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathcal{E}_u(f)$$

Definition 2 (Equality) *The equality measure helps us quantify the fairness of the model by looking at the worst-case performance. The equality of a model f is the maximum error among all users:*

$$\mathcal{E}_{\max}(f) = \max_u \mathcal{E}_u(f)$$

These definitions formalize the trade-off between overall performance and fairness in preference learning models. While traditional machine learning approaches focus on minimizing average error, this can lead to performance disparities across users, potentially perpetuating epistemic injustices. By introducing equality alongside efficiency, we propose a framework for developing models that maintain consistent accuracy across all users.

Our focus on equality, particularly through the $\mathcal{E}_{\max}(f)$ metric, resonates with John Rawls' maximin principle of justice (Rawls 1971; Kenfack, Kahou, and Aïvodji 2024). This approach prioritizes the welfare of the worst-off members, which in our context translates to minimizing the maximum error across all users. This Rawlsian perspective provides a philosophical justification for prioritizing outcomes for disadvantaged users or groups, ensuring that AI systems accurately represent all preferences, including those from marginalized or underrepresented populations.

Equality Metrics

We now introduce a comprehensive set of metrics to quantify the extent to which a model's performance varies across users. These metrics, adapted from the economics literature on income inequality, allow us to measure different aspects of fairness and equality in AI systems.

The importance of these metrics lies in their ability to capture various manifestations of inequality in model performance. For instance, errors might be highly dissimilar across specific groups of users or may vary more gradually across the user population. By employing a range of metrics, each with distinct characteristics, we can gain a nuanced understanding of how fair our preference learning models are. All of the following metrics are non-negative and equal to zero only when the model's performance is identical for all users, representing perfect equality.

Maximal Error Gap The Maximal Error Gap measures the largest discrepancy in model performance between any two users. This metric is particularly useful for identifying extreme cases of inequality and aligns with Rawlsian principles of justice by highlighting the worst-case scenario.

$$G_{\max}(f) = \max_{u,u' \in \mathcal{U}} \left(\mathcal{E}_u(f) - \mathcal{E}_{u'}(f)\right) \tag{1}$$

Standard Deviation of the Error This metric provides a measure of the overall spread of errors across users. A large standard deviation indicates significant variability in model performance, suggesting unequal representation of user preferences.

$$\sigma^2(f) = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \left(\mathcal{E}_u(f) - \bar{\mathcal{E}}(f)\right) \tag{2}$$

Gini Coefficient (Gini 1936) The Gini Coefficient, widely used in economics to measure income inequality, provides a holistic view of error distribution across users. It can be visualized using the Lorenz curve, where the coefficient represents the area between the line of perfect equality and the actual error distribution curve. The Gini Coefficient is bounded between 0 (perfect equality) and 1 (extreme inequality).

$$G(f) = \frac{\sum_{u,u'} \left(|\mathcal{E}_u(f) - \mathcal{E}_{u'}(f)|\right)}{2|\mathcal{U}|^2 \bar{\mathcal{E}}(f)} \tag{3}$$

Generalised Entropy Index (Shorrocks 1980) This index offers flexibility through its α parameter, allowing us to focus on different parts of the accuracy distribution across users. Lower α values are sensitive to the existence of users with low accuracy. While higher α values are more sensitive to the existence of users with high accuracy.

$$G_\alpha(f) = \begin{cases} \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \frac{\mathcal{E}_u(f)}{\bar{\mathcal{E}}(f)} \ln \frac{\mathcal{E}_u(f)}{\bar{\mathcal{E}}(f)} & \text{if } \alpha = 1 \\ -\frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \ln \frac{\mathcal{E}_u(f)}{\bar{\mathcal{E}}(f)} & \text{if } \alpha = 0 \\ \frac{1}{|\mathcal{U}|\alpha(\alpha-1)} \sum_{u \in \mathcal{U}} \left[\left(\frac{\mathcal{E}_u(f)}{\bar{\mathcal{E}}(f)}\right)^\alpha - 1\right] & \text{else} \end{cases} \tag{4}$$

Atkinson Index (Atkinson et al. 1970) Similar to the Generalized Entropy Index, the Atkinson Index uses an ϵ parameter to focus on inequalities at different ends of the accuracy distribution across users. As ϵ increases, the index becomes more sensitive to errors at the lower end of the distribution.

$$A_\epsilon(f) = \begin{cases} 1 - \frac{1}{\bar{\mathcal{E}}(f)} \left(\prod_{u \in \mathcal{U}} \mathcal{E}_u(f)\right)^{1/N} & \text{if } \epsilon = 1 \\ 1 - \frac{1}{\bar{\mathcal{E}}(f)} \min_{u \in \mathcal{U}} \mathcal{E}_u(f) & \text{if } \epsilon = +\infty \\ 1 - \frac{1}{\bar{\mathcal{E}}(f)} \left(\frac{\sum_{u \in \mathcal{U}} \mathcal{E}_u(f)^{1-\epsilon}}{|\mathcal{U}|}\right)^{\frac{1}{1-\epsilon}} & \text{else} \end{cases} \tag{5}$$

Kutznets Ratio (Kuznets 2019) Unlike the previous metrics that consider the entire error distribution, the Kuznets Ratio focuses on the extremes, comparing the errors of the top $\alpha\%$ of users to the bottom $\alpha\%$. This metric is particularly useful for identifying disparities between the best and worst-served users:

$$K_\alpha(f) = \frac{\sum_{\text{top } \alpha\%} \mathcal{E}_u(f)}{\sum_{\text{bottom } \alpha\%} \mathcal{E}_u(f)} \tag{6}$$

By employing this diverse set of metrics, we can comprehensively evaluate the fairness and equality of preference learning models, allowing us to focus on different ends of the distribution.

Experimental Setup

To rigorously evaluate our proposed equality metrics, we carefully selected three datasets that provide crucial individual-level annotation data. This granular information—detailing which user provided each annotation—enables us to precisely analyze variations in model performance across diverse users. Such user-specific data is instrumental in uncovering potential epistemic injustices in AI systems, yet it is often absent from many widely-used datasets in the field of Reinforcement Learning from Human Feedback (RLHF).

Notably, prominent datasets such as Safe RLHF (Dai et al. 2024), Helpful and Harmless (Bai et al. 2022), WebGPT (Nakano et al. 2022), ImageReward (Xu et al. 2023), and AVA (Murray, Marchesotti, and Perronnin 2012) lack annotation at the user level, annotation in those datasets are user agnostic and assume that different users would provide the same annotations. This omission precludes the differentiation of users with diverse preferences, rendering the computation of our proposed equality metrics impossible on these datasets. The absence of such critical information in these widely-used resources highlights a significant gap in the field's ability to assess and address epistemic injustice in generative models.

We posit that the inclusion and release of annotations that include anonymized user IDs, is not merely beneficial but essential for the comprehensive evaluation of fairness in AI systems. By enabling the application of metrics like those proposed in this study, such data would significantly enhance our capacity to identify, quantify, and ultimately mitigate epistemic injustices in generative models. This underscores the urgent need for more nuanced and comprehensive datasets in the pursuit of truly fair and equitable AI systems.

Pick-a-pic Dataset

Dataset We employ our inequality metrics on the Pick-a-pic dataset (Kirstain et al. 2023) for preference learning over generated images. The Pick-a-pic dataset contains over 500,000 examples and over 35,000 prompts. Each example has 2 images generated using Stable-Diffusion for a single prompt by a particular user. The dataset is used for pairwise preference learning tasks over generated images. We chose to evaluate our metrics with this dataset as it is among the few prominent preference learning datasets with user-specific annotation information.

Model We demonstrate our metrics using the open-sourced CLIP-style PickScore model (Kirstain et al. 2023), fine-tuned on the Pick-a-pic V1 dataset to score the quality of the generated image given a prompt. We chose the PickScore model because we want a mainstream utility function that is trained to learn human preference over a particular task. We evaluate the metrics on the test Pick-a-pic V1 dataset to evaluate inequality over both in-distribution users and out-of-distribution users.

AI-EDI-Space Dataset

Dataset The AI-EDI-Space dataset (Gowaikar et al. 2024) is a preference learning dataset that consists of 7,833 street-view images representing a diverse set of public spaces from the Montreal Metropolitan Area. The dataset includes 19,990 pairwise comparisons, evaluated by 22 individuals who were selected to maximize diversity and include underrepresented groups based on ethnicity, gender, sexuality, and age. Each participant evaluated a minimum of 500 comparisons based on 35 different criteria designed to capture various qualities of public spaces. The comparisons contain a real-valued score between -1 and 1, to avoid Arrow's impossibility theorem (Arrow 1950), as explained by (Allouah et al. 2024), but this introduces complexity to the voting patterns, as illustrated in Figure 1.

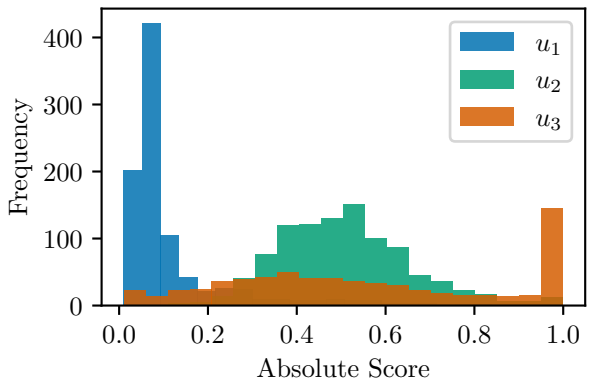


Figure 1: Voting patterns observed in the AI-EDI-Space Dataset. We plot the histogram of the absolute scores given by 3 participants with clearly distinct voting patterns. This figure highlights that users might have very different "taste profiles".

This dataset is specifically chosen to portray the real-world impact of the epistemic injustice caused by diverse human preferences. This dataset is ideal for testing algorithmic equity, as it includes a diverse set of participants with potentially divergent views on what makes a public space valuable, making the task highly subjective.

Model The model and training procedure used are similar to the one proposed in (Dubey et al. 2016). The model takes a single image as input and predicts the scores. The model consists of a feature extractor and a classifier head. The feature extractor is a pre-trained model. We experimented with several models, including VGG11 (Simonyan and Zisserman 2014), EfficientNet (Tan and Le 2019), Squeezenet (Iandola et al. 2016), and DinoV2 (Oquab et al. 2023). The extracted features were then passed through a classifier head to predict a score for each of the 35 criteria. We observed that the model with EfficientNet, along with a double-layered classification head with 256 as the hidden dimension, gave the best results throughout all the experiments. Hence, all values for the AI-EDI-Space dataset are using an EfficientNet feature extractor. To train the model, we computed the scores for both images in each comparison and calculated the difference in scores between the two images. We then used the Mean Squared Error between the difference in scores and the ground truth score as the error to train the model.

Loss We use the 0-1 loss as the baseline metric to compute the various equality metrics, which corresponds to measuring the difference in accuracy across users.

Jester Jokes Dataset

Dataset We also test the proposed metrics on the Jester Jokes Dataset, originally developed for recommender systems research (Goldberg et al. 2001). This dataset contains 100 jokes, each rated on a scale from -10 to +10 by 73,421 participants. The inherently subjective nature of humor appreciation makes this dataset particularly suitable for examining model performance across diverse user preferences. To ensure data quality and diversity, we only selected a subset of the annotations. We first filtered the dataset to include only participants who rated all 100 jokes, ensuring comprehensive engagement from each user. From this filtered set, we then selected 1,000 users exhibiting the most diverse voting patterns. This selection was achieved through Principal Component Analysis (PCA) of users' score vectors, followed by a uniform sampling of users maximally distant in the PCA space. This careful curation process resulted in a dataset that not only maintains balance during model training but also captures a wide spectrum of user preferences.

Model Since users directly provided scores for each joke, the problem can be formulated as a regression task. We trained a model that takes a single joke as input and predicts its score. The model consists of a feature extractor and a classifier head. The feature extractor is a pre-trained BART model (Lewis et al. 2020). The extracted features are then passed through a classifier head, which is either a single-layer or a double-layer perceptron, to predict the score. The Mean Squared Error (MSE) between the predicted score and the ground truth score was used as the loss function to train the model.

Loss We used the MSE loss as the baseline metric to compute the various equality metrics.

Methods

To address the challenge of inequality in model performance across users, we propose and evaluate two categories of techniques: pre-processing and in-processing. These approaches aim to enhance the fairness of preference learning models by ensuring more equitable representation of diverse user preferences.

Pre-Processing Techniques

Pre-processing techniques are applied to the data prior to model training. We investigate three scaling methods designed to normalize the distribution of scores across users:

1. **Min-Max Scaling**: This technique scales each participant's scores to a range of [-1, 1]. It is applied individually to each user's scores, preserving relative preferences within a user's data while enabling comparability across users.
2. **Normalization Scaling**: This two-step process first applies standard normalization to each participant's scores, adjusting the mean to 0 and standard deviation to 1. Subsequently, the scores are scaled to ensure they remain within the [-1, 1] range. While this method, like Min-Max Scaling, is user-specific, it does not guarantee sparse unanimity.
3. **Mehestan Scaling (Allouah et al. 2024)**: This more sophisticated approach considers the voting patterns of all participants when scaling an individual's scores. The process involves: a) Converting raw comparison scores to individual scores using a Generalized Bradley-Terry Model (Fageot et al. 2024). b) Scaling and translating these scores using the BrMean primitive, which is designed to be resilient to potential manipulation by malicious voters. c) Preserving individual score distributions without final aggregation, maintaining the uniqueness of each participant's preference pattern.

Mehestan Scaling is particularly effective in achieving sparse unanimity, a property that ensures the preservation of unanimous preferences even when user voting patterns differ significantly.

In-Processing Techniques

In-processing techniques are integrated into the model training process. We explore two primary approaches:

1. **User Embeddings**: By incorporating user-specific embeddings as additional input to the model, we aim to capture and adapt to individual voting patterns. This technique allows the model to learn user-specific features that may influence preference judgments.
2. **Contrastive Loss**: We employ contrastive loss in conjunction with least squares error (LSE). The contrastive

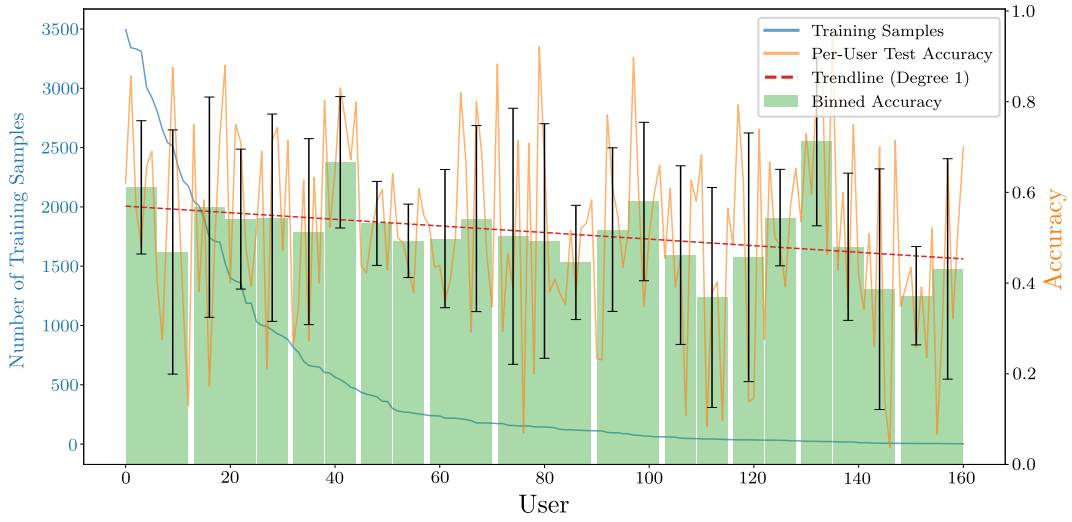


Figure 2: The Training Samples vs Per-User accuracy over the test set of the Pick-a-pic V1 Dataset. No clear accuracy trend can be seen as the training samples decline

loss works by increasing the distance between dissimilar scores, ensuring that the model's outputs are not clustered too closely together. This helps prevent comparisons—calculated as the difference between the scores of two alternatives—from being too close to zero, thereby improving the model's ability to distinguish between different preferences.

Results

Pick-a-pic Dataset We compute the metrics over the Pick-a-pic V1 dataset using the OpenAI CLIP model (Radford et al. 2021) as the baseline. We compare these baseline results to PickScore (Kirstain et al. 2023), a variant of a CLIP-based model fine-tuned on the Pick-a-pic V1 dataset. We compare the metrics for all users as well as users with more than 20 test samples for statistical significance. While sampling the users, we do not put any conditions over the number of training samples as we only observe a weak formative trend in test-accuracies in comparison to the training samples (Figure 2).

From results in Table 1 we can observe a large maximal accuracy that implies a large range in the per-user accuracies. Also, a large standard deviation of per-user accuracies implies a varied distribution of per-user accuracies, suggesting the presence of epistemic injustice. Additionally, a smaller Gini Coefficient (closer to 0) implies a lower *average* inequality, but a large Atkinson Index (closer to 1) implies a larger *extreme* inequality. This suggests inequality in the extremes of the per-user accuracy distribution. This is because the Gini Coefficient works over the entire range, whereas the Atkinson Index with a high α is more susceptible to inequality within the lower part of the distribution. This is further reinforced by a high Kutznets ratio which compares performance disparity within the extremes. We can also observe this behavior in Figure 3, where the Lorenz

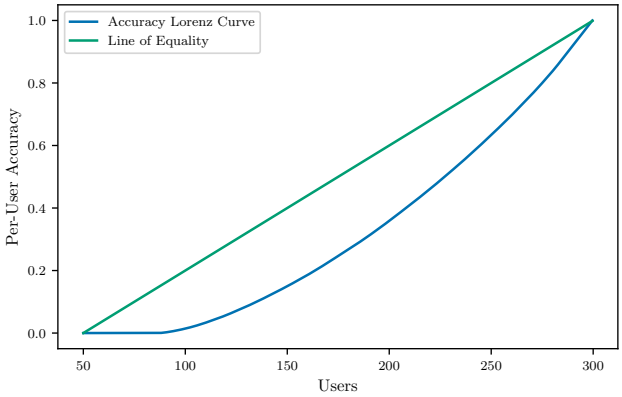


Figure 3: The Lorenz Curve of Per-User Accuracy and the Line of Equality over the Pick-a-pic V1 testset. The Lorenz curve is far from the line of equality at the extremes and makes up over the means

Curve (Fellman 2011) of the Per-User accuracy, indicative of the cumulative per-user accuracy, diverges for users with lower accuracy.

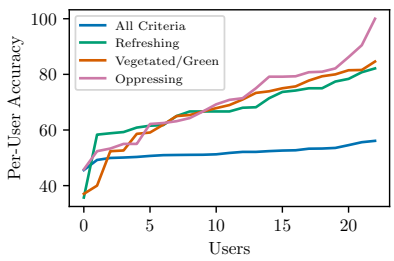
Lastly, epistemic injustice is reflected in the lower inequality metric values for out-of-distribution users compared to in-distribution users, also highlighting the trade-off between utility and equality.

This scenario also highlights the need for multiple metrics to comprehensively understand the inequality, as each metric caters to a particular part of the distribution.

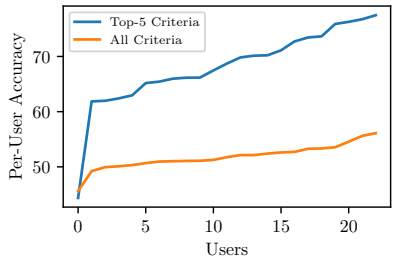
AI-EDI-Space Dataset After training a model on the AI-EDI-Space dataset, we computed the various equality metrics proposed earlier. Figure 4 shows the distribution of accuracy for all users across different criteria. Our analysis

Experiment	Accuracy ↑	G_{max} ↓	σ^2 ↓	G ↓	$G_{\alpha=0}$ ↓	$A_{\epsilon=\infty}$ ↓	$K_{\alpha=20}$ ↓
In-Distribution Users							
CLIP vit-base	46.90	100	29.10	0.18	3.74	1.00	28.09
PickScore	59.97	100	29.30	0.14	3.43	1	8.13
CLIP vit-base^{20+}	51.25	91.95	20.14	0.11	2.62	0.93	3.51
PickScore^{20+}	65.40	92.00	18.46	0.08	2.52	0.95	2.38
Out-of-Distribution Users							
CLIP vit-base	51.27	100	31.30	0.17	2.87	1.00	26.22
PickScore	59.99	100	30.44	0.14	2.47	1	9.31
CLIP vit-base^{20+}	50.72	72.67	18.57	0.10	1.60	0.74	3.22
PickScore^{20+}	64.22	78.71	17.11	0.07	1.49	0.79	2.12

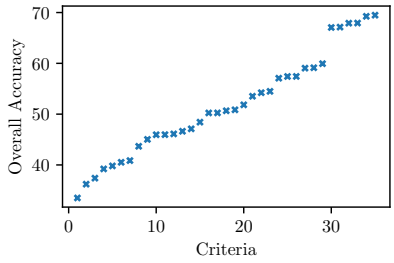
Table 1: The results presented are from the Pick-a-pic V1 Dataset test set where results were computed using individual per-user accuracy. Experiments with a superscript (20+) denote evaluations on a subset of the users with at least more than 20 samples. For each metrics, ↑ indicates that higher values are better while ↓ indicates that lower values are better. For all the metrics a value of 0 represents a situation of perfect equity, where all users have identical accuracy, except for Kuznets ratios K_α for which the optimal value is 1. The Gini Coefficient G and Atkinson Index $A_{\epsilon=100}$ have a maximum value of 1, signifying substantial inequality. In particular, an Atkinson Index of 1 indicates that at least one user has zero accuracy, highlighting significant performance disparities.



(a) Sorted Per User Accuracy over i) all Criteria, ii) Top 3 Criteria — Oppressing, Vegetated/Green, Refreshing (in descending order of accuracy).



(b) Sorted Per User Accuracy over i) All Criteria, ii) only 5 Criteria with highest overall accuracies (viz. Intimate, Regenerative, Refreshing, Vegetated/Green, Oppressing).



(c) Scatter plot of accuracies for every criterion. Highest: 'Oppressing'; Lowest: 'Inviting/Welcoming'.

Figure 4: Experiment results on the AI-EDI Space Dataset. (a) Trendline of per-user accuracy over all and top-3 criteria. (b) Accuracy for all vs. top-5 criteria. (c) Scatter plot of accuracy per criterion.

yielded several noteworthy observations: inequality appears to be higher for criteria where the model performs best overall, as shown in Table 2. This illustrates the potential trade-off between efficiency and equality. The observed inequality seems to be primarily driven by voting patterns, as the mean squared error (MSE) used to train the model penalizes discrepancies with the user's comparisons. We also observe that users whose voting patterns cluster around 0 (a conservative voting approach) tend to achieve higher accuracy. Additionally, the number of comparisons annotated by each user may differ, potentially contributing to the observed inequalities. However, disentangling these effects requires further analysis to understand the influence of sample size on our observations.

Jester Jokes Dataset Our analysis of the Jester Jokes dataset provided additional insights into the effectiveness of various scaling and translation techniques in addressing inequality. Table 3 presents the results of our equality metrics for different approaches. We made the following observatoins: 1) Normalization and MinMax Scaling achieved low Mean Squared Error (MSE) but failed to significantly improve equality. Notably, the Kuznets ratio remained close to 4, indicating substantial inequality between the top 20% and bottom 20% of participants in terms of model performance. 2) Mehestan Scaling, designed for sparse unanimity (Allouah et al. 2024), yielded higher equality values despite a worse MSE. This finding suggests that techniques prioritizing unanimous preference recovery may contribute to greater epistemic fairness. 3) Given our careful selection of users who scored all 100 jokes, we can primarily attribute the observed inequity to differences in voting patterns rather than data imbalance. This reinforces the importance of considering diverse preference expressions in model development. Additionally, we can clearly see the tradeoff between performance and equality from Figure 5 where the aver-

Experiment	Accuracy ↑	G_{max} ↓	σ^2 ↓	G ↓ $\times 10^{-2}$	$G_{\alpha=0}$ ↓ $\times 10^{-4}$	$A_{\epsilon=\infty}$ ↓ $\times 10^{-1}$	$K_{\alpha=20}$ ↓
Normalisation	51.1 ± 1.9	6.6 ± 1.1	1.7 ± 0.2	0.9 ± 0.1	5.5 ± 1.5	0.7 ± 0.1	1.37 ± 0.02
MinMax	50.2 ± 2.6	7.6 ± 4.0	1.8 ± 0.8	1.0 ± 0.4	7.1 ± 5.8	0.7 ± 0.6	1.38 ± 0.03
Mehestan	51.3 ± 4.4	7.1 ± 1.2	1.8 ± 0.3	1.1 ± 0.2	7.4 ± 3.4	0.7 ± 0.2	1.39 ± 0.04
Contrastive Loss	52.0 ± 0.6	7.7 ± 1.8	2.1 ± 0.6	1.1 ± 0.3	8.0 ± 4.7	0.8 ± 0.1	1.40 ± 0.04
User Emb.	50.2 ± 0.7	7.8 ± 1.0	1.9 ± 0.4	1.0 ± 0.2	7.0 ± 2.6	0.8 ± 0.4	1.39 ± 0.03

Table 2: Results for the AI-EDI Image Dataset over different experiments. Individual Users evaluated using Per-User Accuracy

Experiment	MSE ↓	G_{max} ↓	σ^2 ↓	G ↓	$G_{\alpha=0}$ ↓	$A_{\epsilon=\infty}$ ↓	$K_{\alpha=20}$ ↓
Normalisation	0.27	0.76	0.14	0.14	0.16	9.6 ± 0.1	4.81 ± 0.08
MinMax	0.26	0.78	0.14	0.14	0.16	9.6 ± 0.3	4.84 ± 0.04
Mehestan	0.89	2.19	0.34	0.10	0.07	6.7 ± 0.4	2.81 ± 0.37

Table 3: Results for the Jester Jokes Dataset over different experiments. Individual Users evaluated using Per-User Mean Squared Error

age MSE falls over training epochs, indicating performance gains, whereas the Gini coefficient and Maximal MSE increase, indicating increased inequality.

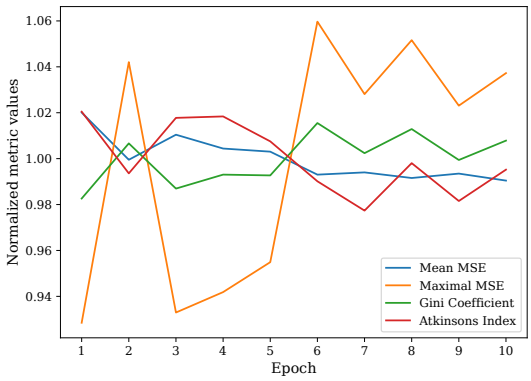


Figure 5: Performance and Metrics for Jester Jokes Dataset over epochs. A clear tradeoff in performance and equality can be seen from the figure where equality seems to increase over training epochs and performance decreases. A trendline of accuracy and equity metrics over training epochs showing a clear inverse relation between equity and training progress

Conclusion

In the era of generative AI, where algorithms increasingly shape decision-making processes, ensuring that these systems do not generate or amplify epistemic injustice is paramount. This study introduces a novel perspective on fairness in preference learning, focusing on the equitable representation of diverse human preferences and views. Our work provides a framework for quantifying and addressing epistemic fairness in AI models, contributing to the development of more just and inclusive AI technologies. Our findings underscore the potential for significant disparities in how well AI models capture preferences across different users. This raises critical questions about epistemic justice in AI systems and highlights the need for further research in several key areas: 1) Further investigation is needed to understand the sources of inequality in model performance and develop effective mitigation strategies. This includes examining how data characteristics, model architectures, and diverse human preferences interact to produce or exacerbate inequalities. 2) As Reinforcement Learning from Human Feedback (RLHF) becomes more prevalent, ensuring that the alignment process itself is equitable across diverse participant groups is crucial. This involves developing methods to capture a wide range of opinions and preferences, particularly from marginalized or underrepresented groups. 3) We advocate for the public release of more datasets that include annotations that includes anonymized information about annotator, detailing which annotator provided which annotation. This granular data is crucial for conducting comprehensive evaluations of epistemic justice in AI models. Such datasets would enable researchers to track how different users' preferences and judgments are represented in model outputs, providing a more nuanced understanding of potential biases or inequalities in preference learning and generative AI systems. Future work should explore how to balance these concerns in line with principles of distributive justice and epistemic fairness, particularly in the context of generative AI systems.

This study represents a step towards more equitable AI systems that respect and accurately represent diverse human preferences. As AI continues to play an increasingly significant role in society, addressing epistemic fairness will be crucial in ensuring that these systems serve all members of society equitably. Our work provides a foundation for future research in this critical area, aiming to develop AI technologies that are not only efficient but also just and inclusive.

References

Allouah, Y.; Guerraoui, R.; Hoang, L.-N.; and Villemaud, O. 2024. Robust sparse voting. In *International Conference on Artificial Intelligence and Statistics*, 991–999. PMLR.

Aroyo, L.; and Welty, C. 2015. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Magazine*, 36(1): 15–24.

Arrow, K. J. 1950. A Difficulty in the Concept of Social Welfare. *Journal of Political Economy*, 58(4): 328–346.

Atkinson, A. B.; et al. 1970. On the measurement of inequality. *Journal of economic theory*, 2(3): 244–263.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; Joseph, N.; Kadavath, S.; Kernion, J.; Conerly, T.; El-Showk, S.; Elhage, N.; Hatfield-Dodds, Z.; Hernandez, D.; Hume, T.; Johnston, S.; Kravec, S.; Lovitt, L.; Nanda, N.; Olsson, C.; Amodei, D.; Brown, T.; Clark, J.; McCandlish, S.; Olah, C.; Mann, B.; and Kaplan, J. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv:2204.05862.

Bakker, M. A.; Chadwick, M. J.; Sheahan, H.; Tessler, M. H.; Campbell-Gillingham, L.; Balaguer, J.; McAleese, N.; Glaese, A.; Aslanides, J.; Botvinick, M.; and Summerfield, C. 2022. Fine-tuning language models to find agreement among humans with diverse preferences. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Beutel, A.; Chen, J.; Doshi, T.; Qian, H.; Wei, L.; Wu, Y.; Heldt, L.; Zhao, Z.; Hong, L.; Chi, E. H.; and Goodrow, C. 2019. Fairness in Recommendation Ranking through Pairwise Comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2212–2220. Anchorage AK USA: ACM. ISBN 978-1-4503-6201-6.

Chakraborty, S.; Qiu, J.; Yuan, H.; Koppel, A.; Huang, F.; Manocha, D.; Bedi, A.; and Wang, M. 2024. MaxMin-RLHF: Towards Equitable Alignment of Large Language Models with Diverse Human Preferences. In *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*.

Cotter, A.; Jiang, H.; Gupta, M.; Wang, S.; Narayan, T.; You, S.; and Sridharan, K. 2019. Optimization with Non-Differentiable Constraints with Applications to Fairness, Recall, Churn, and Other Goals. *Journal of Machine Learning Research*, 20(172): 1–59.

Dai, J.; Pan, X.; Sun, R.; Ji, J.; Xu, X.; Liu, M.; Wang, Y.; and Yang, Y. 2024. Safe RLHF: Safe Reinforcement Learning from Human Feedback. In *The Twelfth International Conference on Learning Representations*.

Dubey, A.; Naik, N.; Parikh, D.; Raskar, R.; and Hidalgo, C. A. 2016. Deep Learning the City: Quantifying Urban Perception at a Global Scale. In Leibe, B.; Matas, J.; Sebe, N.; and Welling, M., eds., *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, 196–212. Cham: Springer International Publishing. ISBN 978-3-319-46448-0.

Fageot, J.; Farhadkhani, S.; Hoang, L.-N.; and Villemaud, O. 2024. Generalized Bradley-Terry Models for Score Estimation from Paired Comparisons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 20379–20386.

Fellman, J. 2011. *Lorenz Curve*, 760–762. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-04898-2.

Fricker, M. 2007. *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

Gini, C. 1936. On the measure of concentration with special reference to income and statistics, Colorado College Publication. *General series*, 208(1).

Goldberg, K.; Roeder, T.; Gupta, D.; and Perkins, C. 2001. Eigentaste: A Constant Time Collaborative Filtering Algorithm. *Information Retrieval*, 4(2): 133–151.

Gowaikar, S.; Berard, H.; Mushkani, R.; Marchand, E. B.; Ammar, T.; and Koseki, S. 2024. AI-EDI-SPACE: A Co-designed Dataset for Evaluating the Quality of Public Spaces. arXiv:2411.00956.

Hardt, M.; Price, E.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, volume 29.

Iandola, F. N.; Han, S.; Moskewicz, M. W.; Ashraf, K.; Dally, W. J.; and Keutzer, K. 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and¡ 0.5 MB model size. *arXiv preprint arXiv:1602.07360*.

Kay, J.; Kasirzadeh, A.; and Mohamed, S. 2024. Epistemic Injustice in Generative AI. *arXiv preprint arXiv:2408.11441*.

Kenfack, P. J.; Kahou, S. E.; and Aïvodji, U. 2024. A Survey on Fairness Without Demographics. *Transactions on Machine Learning Research*.

Kirstain, Y.; Polyak, A.; Singer, U.; Matiana, S.; Penna, J.; and Levy, O. 2023. Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Kovač, G.; Sawayama, M.; Portelas, R.; Colas, C.; Dominey, P. F.; and Oudeyer, P.-Y. 2023. Large Language Models as Superpositions of Cultural Perspectives. arXiv:2307.07870.

Kuznets, S. 2019. Economic growth and income inequality. In *The gap between rich and poor*, 25–37. Routledge.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880. Online: Association for Computational Linguistics.

Mill, J. S. 1859. *On liberty*. Cambridge University Press.

Murray, N.; Marchesotti, L.; and Perronnin, F. 2012. AVA: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2408–2415.

Nakano, R.; Hilton, J.; Balaji, S.; Wu, J.; Ouyang, L.; Kim, C.; Hesse, C.; Jain, S.; Kosaraju, V.; Saunders, W.; Jiang, X.; Cobbe, K.; Eloundou, T.; Krueger, G.; Button, K.;

Knight, M.; Chess, B.; and Schulman, J. 2022. WebGPT: Browser-assisted question-answering with human feedback. arXiv:2112.09332.

Narasimhan, H.; Cotter, A.; Gupta, M.; and Wang, S. 2020. Pairwise fairness for ranking and regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 5248–5255.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Ramé, A.; Couairon, G.; Shukor, M.; Dancette, C.; Gaya, J.-B.; Soulier, L.; and Cord, M. 2023. Rewarded soups: towards Pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. arXiv:2306.04488.

Rawls, J. 1971. *A theory of justice*. Belknap Press of Harvard University Press.

Sandri, M.; Leonardelli, E.; Tonelli, S.; and Jezek, E. 2023. Why Don't You Do It Right? Analysing Annotators' Disagreement in Subjective Tasks. In Vlachos, A.; and Augenstein, I., eds., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2428–2441. Dubrovnik, Croatia: Association for Computational Linguistics.

Sanyal, A.; Hu, Y.; and Yang, F. 2022. How Unfair Is Private Learning ? In *The 38th Conference on Uncertainty in Artificial Intelligence*.

Sawyer, R. L.; Cole, N. S.; and Cole, J. W. L. 1976. Utilities and the Issue of Fairness in a Decision Theoretic Model for Selection. *Journal of Educational Measurement*, 13(1): 59–76.

Shorrocks, A. F. 1980. The class of additively decomposable inequality measures. *Econometrica: Journal of the Econometric Society*, 613–625.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 6105–6114. PMLR.

Xu, J.; Liu, X.; Wu, Y.; Tong, Y.; Li, Q.; Ding, M.; Tang, J.; and Dong, Y. 2023. ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation. arXiv:2304.05977.

Yu, T.; Lin, T.-E.; Wu, Y.; Yang, M.; Huang, F.; and Li, Y. 2023. Constructive Large Language Models Alignment with Diverse Feedback. arXiv:2310.06450.

Zeng, D.; Dai, Y.; Cheng, P.; Wang, L.; Hu, T.; Chen, W.; Du, N.; and Xu, Z. 2024. On Diversified Preferences of Large Language Model Alignment. arXiv:2312.07401.

Ziegler, D. M.; Stiennon, N.; Wu, J.; Brown, T. B.; Radford, A.; Amodei, D.; Christiano, P.; and Irving, G. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.